一种综合性论文查重评估方法 *

谢兆贤，叶淑祯[†]，黄沈权

(温州大学 机电工程学院，浙江 温州 325035)

摘　要：论文查重对高校的毕业论文质量管控的过程是有其必要性的。提出一种新型综合性作法，针对互联网的查重系统过程过于极端的状况下，可以使用此方法检测出异常，才会要求再次以人工检测。目标是要减少误判，将原本不是抄袭却被判定抄袭的论文得到申覆的机会。当同一篇论文在不同的查重网站中检测出的重复率相差较大时，采用论文新型综合性查重方法，在判定论文最后的重复结果中加入人为判断的权重，降低论文重复率被网站所控制的因素，使得查重结果不完全被网站或人工控制，这种网站人工的双重混合式检测弥补了网站资源库问题对查重结果产生的影响，提升论文查重结果的准确性和可信度。

关键词：毕业论文；查重网站；人工检测；剽窃

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.03.0156

Comprehensive weight checking method for graduation thesis

Xie Zhaoxian, Ye Shuzhen[†], Huang Shenquan

(*School of mechanical & electrical engineering*, *Wenzhou University*, *Wenzhou Zhejiang* 325035.)

Abstract: It is necessary for the quality control process of university graduation thesis to check papers repetition rate. Due to the requirements of detecting repetition rate, most of them can only do the similar proportion and quantity simply by manual examination or by the Internet. In this paper, it proposes a new comprehensive approach including the manual detection because the related Internet system is not fully trustable. The goal is to reduce miscarriage of justice and get the chance to submit papers that were not copied and plagiarized. When the repetition rate of the same papers which is detected in different sites has a huge difference, a comprehensive weight checking method is adopted in this paper. This paper adds artificial judgments to the repeated results and reduces the impact of the website. As a result, the checking results are not completely controlled by website or manual work. The website and artificial double hybrid detection of papers makes up the influence of the website resource library on the result of checking the weight, and improves the accuracy and credibility of the checking result.

Key words: graduation thesis; duplicate checking website; manual detection; plagiarism

0　引言

目前国内的许多高校在学生毕业之前都会要求学生撰写毕业论文，目的乃在培养学生在此过程得到整体知识的理解、验证、和写作的能力。然而，尽管指导老师很认真地提供题目与指导学生，对毕业论文内容的掌控上，仍然无法完整巨细靡遗地检验每位学生的研究成果。也就是说，面对学生论文抄袭的部分，仍是有所欠缺[1]。所幸在互联网[2]发达的今天，许多学校为了确保客观地评分都会先交给查重网站，让一些网站做先期的筛选以便取得论文相似性的检验[3,4]。

故此，衍生一些商机与问题。商机部分，许多查重网站的建立，有免费的查重网站[5~7]与付费的查重网站[8~15]，主要提供作者或是相关机构对特定论文内容做审查[16]。产生的问题说明如下，尽管查重网站对整体教育做到帮补的功用，从相似性上可以挑选出疑似抄袭的论文，依照不同网站所收藏的数据库内容决定相似性的百分比。此时，在不同查重机制下所产生的重复比例不见得相同，重复比例高的也不能确定它的真实性[17]。换句话说，传统查重网站的目的倾向找出相同文字的功能，无法防止或误认的状况。所以，本论文强调降低误认的作法，将针对既有查重方式进行分析与建议，进而提出一种新的论文查重流程与架构。

1　问题与设计

一般情况下，检查抄袭的过程，可以分成几种状况:真的抄袭被找出;没有抄袭被找出;真的抄袭没被找出，这种状况代表查重网站的数据库量不足或是被抄袭的对象没有被收纳入查重

收稿日期：2018-03-11；**修回日期：**2018-05-12　**基金项目：**国家自然科学基金资助项目（71501143）；浙江省自然科学基金资助项目（LQ14G010006）

作者简介：谢兆贤（1971-），男，副教授，博士研究生，主要研究方向为云计算、软件工程、智能制造（george_hsieh@qq.com）；叶淑祯（1996-），女（通信作者），本科，主要研究方向为智能制造、物联网；黄沈权（1986-），男，讲师，博士，主要研究方向为计算机集成制造、知识工程.

网站;没有抄袭也没被找出。根据以上四种状况，第一类状况正是所期待的结果，文章的作者确实有抄袭也被找出，如此便能帮补指导老师或是评审的失误。第三类受限于特定数据库内部文章或是该网站的算法涵盖面低，这种状况属于网站本身的问题，是有可能发生。第四类状态也是所期待的结果，没有抄袭也没被找出。然而第二类的情况才是所关心的，文章作者本身没有抄袭却被评出高相似性，最后，被误判为抄袭。

除此之外，由于每个查重网站所使用的算法与数据库都不相同，查重结果比例也就不同[18]。那么该听谁的呢?查重结果比例越高越好吗? 这个是我们质疑的，因为只要以较严格的算法那么自然就可以提升重复率。总结以上内容，此篇论文将要解决两个问题:针对没有抄袭被找出的状况来给予协助;查重比例多少才是可靠的问题。

对于以上的问题，本文提出的解决方法有如下几点：a)分析数种方法修改参考文献与文章，使之降低查重的误判机会,如修改数学符号与文字描述方式;b)利用统计方法与归纳法分析具可靠性的查重比例。对产生的样本文章，针对目前的查重网站先行分析差异后，输入免费查重网站并且取得输出结果做多次实验。设计多种状态与实验方法，最后呈现整体的关系并提出结论。

2　系统架构与方法

2.1　国内查重网站的分类和描述

针对国内目前使用率较高的查重网站,通过实际测试总结出了各个查重网站的优缺点：a）知网[8]的查重准确度高，查询速度快，但是查询价格偏高；b）维普网[9]检测费用少、检测速度快、准确度较高，但是不能识别论文中的表格，其他外语的论文收录非常地少，使用不方便；c）万方检测网[10]检测费用低，但是资源量较少；d）PaperPass 论文查重网[11]可实时在线修改论文，重复点比对效率高，准确率较高，但是不能检测英文，准确性中等；e）知识产权卫士-拷克网[12]提供许多网页数据基础，能检测论文的相似度，提供抄袭检测报告，支持英文检测，但是资源量不是很多，查重准确性中等；f）中国搜文章照妖镜[6]免费提供检测，可快速进行检测和判定，且可以检测抄袭量，但是该检测网站功能较差，检测不稳定，准确性较低；g）大雅[5] 免费提供检测，且每天查询不限次数，查询速度快，但是其资料库有限，查询结果仅供参考；h）格子达[7] 免费提供检测，检测结果比较精准，有在线修改功能，但是不能检测英文，资源库不全面，查询结果中引用率过高抄袭率偏低；i）PaperFree[13] 可以准确地查到论文中的潜在抄袭和不当引用，可以边修改边检测，改哪里检测那里，按实际修改句子收费，不改的内容不收费，但是其查重不是非常严格；j）论文狗[14] 不限字数纯免费使用，学术期刊资源库和互联网实时更新资源，但是数据资源较少，查重不是非常准确；k）PaperTest[15]检测速度快，还提供修改建议，但是数据文献不是特别广泛。

2.2　系统架构

如图 1 所示，进入查重网站查重的一般过程为首先在查重论文输入框中输入数据形态与做法，然后查重系统按照一定的算法处理论文，最后输出数据形态与做法，即输出查重结果，包含重复率部分和重复的来源等信息。整体过程可以简略成输入、处理方式和输出三部分。

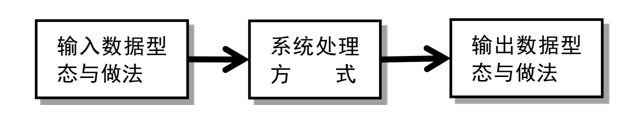


图 1　系统处理基本流程

图 2 为论文投入免费查重网站获得查重结果的实验方法图。本文的做法是将同一篇论文分别输入 A、B、C 三个系统处理，分别得到对应的查重数据结果 a、b、和 c，将查重数据结果 a、b、和 c 进行比较，发现查重规律和查重问题，得出结论，提出解决方法。

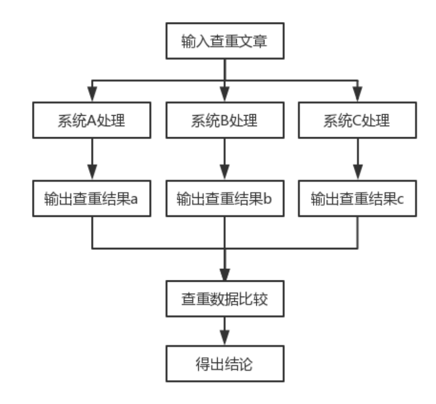


图 2　查重实验方法图

基于目前查重的架构下，本文发展一个改进版的查重架构，简称新型查重(new check repeat，NCP)的架构，如图 3 所示。该整体架构包括以下五个方面：

a)采集层。采集层中包括一个输入模块。用户通过输入模块将需要查重的稿件输入，作为系统采集信息的入口。

b)数据层。数据层包括数据库和全文检索系统。数据库是稿件查重应用领域的通用数据处理系统，是组织、存储和管理数据的仓库；全文检索是计算机程序通过扫描文章中的每一个词，对每一个词建立一个索引，指明该词在文章中出现的次数和位置，当用户查询时根据建立的索引查找，类似于通过字典的检索字表查字的过程，全文检索系统是按照全文检索理论建立起来的，将用于提供全文检索服务的软件系统。

c)处理层。处理层包括一个查询计算模块。查询计算模块是查重网站根据一定的算法计算出稿件的重复率，进行稿件抄袭检查。通过该模块可得到稿件重复部分和重复的来源等信息。

d)核心业务层。核心业务层包括测试模块和人工判断模块。测试模块是将同一篇稿件投入不同的查重网站得到的重复率进行测试比对，并判断查重结果；人工判 断模块是将稿件进行人工复合审查，判断重复率。

e)用户。新型查重系统服务的用户包括新闻出版管理单位、科研管理单位、杂志社、大学和公众等。

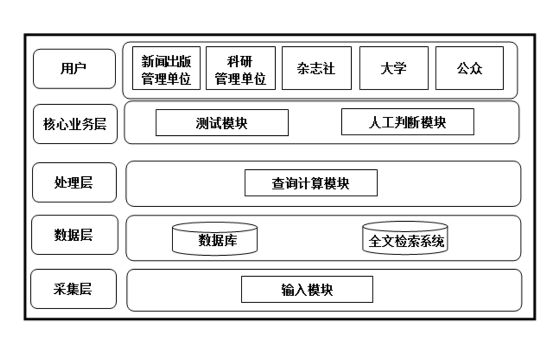


图 3　新型查重架构图

2.3　方法

经过查重实验的分析，得出一个新的查重流程图（如图 4 所示）来提高论文查重的准确性，尽量防止误判情况的发生。

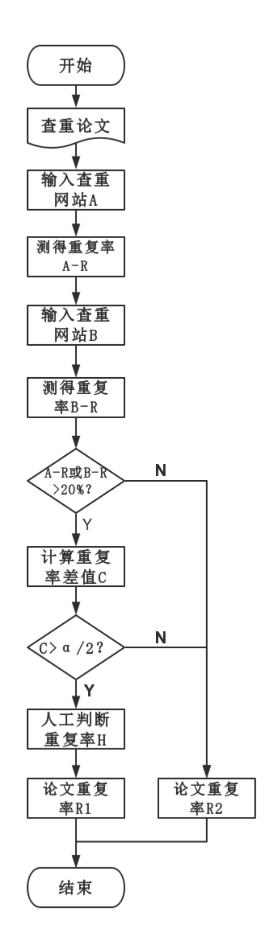


图 4　查重流程图

具体的查重流程为：先通过输入模块将需要查重的论文输入查重网站 A，再把论文输入查重网站 B，通过查询计算模块进行数据库和全文检索可得网站 A 测得该论文的重复率为 A-R，网站 B 测得该论文的重复率为 B-R。然后测试模块进行判断，如果 A-R 和 B-R 都小于 20%（中国鸣网在期刊论文百科问答中提到"抄袭过多，一经查出超过 20%，后果严重"[19]，因此这里取值 20%作为判断依据），表明两个网站一致认为该论文重复率低，则判定该论文的重复率为 R2 如下所示。

$$R2=\alpha\times\max\{A-R,B-R\}+(1-\alpha)\times\min\{A-R,B-R\} \quad (1)$$

其中：α 是置信度，为一个事先给定值，一般来说，取对称区间时，区间长度越小，越准确；置信度取得越高，区间估计的可信程度越高，但区间估计的精确度会降低[20]。这里 α 的值由毕业论文审查人根据论文实际情况决定，通常根据统计学取 95%的置信度[21]。在计算 R2 值时，因为数值较大的查重结果意味着该网站收录有更多与输入论文有关的资源，对查重结果影响较大，可信度较高，所以把较大的权重 α 赋给较大的重复率，考虑到另一个网站对查重结果也有一定影响，所以把较小的权重（1-α）赋给较小的重复率[22]。

否则，计算 A-R 和 B-R 间的差值 C，如果 C 小于 α/2，表明两个网站的查重结果误差在允许范围之内，则判定该论文的重复率为 R2；否则，加入人工判断该论文的重复率 H 作为参考，则判定该论文的重复率为 R1，如下所示。

$$R1=80\%\times H+20\%\times\alpha\times\max\{A-R,B-R\}+20\%\times(1-\alpha)\times\min\{A-R,B-R\} \quad (2)$$

式中的 α 的意义同式（1）。在计算 R1 值时，因为重复率不能超过 20%，所以把 80%的权重赋值给人工判定的重复率 H，其意义在于即使网站判定论文为完全抄袭，只要审核该论文的专家根据经验确定其为未抄袭论文，该论文仍可被判定为未抄袭。

图 4 的查重流程范例如下。先将需要查重的论文输入查重网站 A，网站 A 测得该论文的重复率 A-R=15%，再把论文输入查重网站 B，网站 B 测得该论文的重复率 B-R=80%。由于 B-R=80%>20%，计算重复率差值 C=|(A-R)-(B-R)|=|15%-80%|=65%，C=65%>α/2=47.5%（此处 α 取 95%），然后进行人工判断。人工判断重复率 H=18%，则该论文重复为

$$\begin{aligned}R1&=80\%\times H+20\%\times\alpha\times\max\{A-R,B-R\}+20\%\times(1-\alpha)\times\min\{A-R,B-R\}\\&=80\%\times18\%+20\%\times95\%\times80\%+20\%\times(1-95\%)\times15\%\\&=29.75\%>20\%\end{aligned}$$

结果可得该论文重复率较高，判定为抄袭论文性质。

3　实验与分析

3.1　实验内容

现取一篇名为《帮助高中生渡过数学学习困难期的几点尝试》[23]的论文，对论文做四种改动，只改公式、有意义地改文字、无意义地改文字和改排版。

a)只改公式-改公式前和改公式后。改公式是指改变论文中公式的数字、运算符号、算法、函数、逻辑等方面的操作。

b)只改文字-改文字前和改文字后。修改文字后会有两种状况，一种是有意义的文字，另一种是无意义的文字。有意义地改文字是指通过有弹性地更改论文中的某些词组或句子使得文段的原意改变为另一种意思的操作。无意义地改文字是指通过机械式地更改论文中的某些词组或句子使得文段变得无意义的操作。

c)改排版-改文字排版前和改文字排版后。改排版是指将文字、图片、图形等可视化信息元素在版面布局上调整位置、大小的操作。实施此案例之后，或许会降低重复率，但是整句还是类似但不易阅读与理解，容易造成查重的误判，这不是此篇论文所关心的议题。此外，发展新的查重算法也不是此篇论文

要讨论的。

3.2 分析结果

此篇论文检查论文重复率所使用的网站是大雅[5]、中国搜文章照妖镜[6]和格子达[7]，这三个免费论文查重网站是目前使用较为广泛的。

图 5 中所提及的标准化（Normalization），是指对重复率做数据标准化处理。即在某一查重网站中，以测得的原文重复率为标准，计算出论文改动后测得的重复率与原文重复率之比。在四种原文改动方式下，即只改公式、有意义地改动文字、无意义地改动文字和只改排版。图 5 画出了三个查重网站重复率标准化后的柱状图。此外，可以发现将文章只改排版，中国搜文章照妖镜标准化后的数值为 1.6，即只改排版后中国搜文章照妖镜测得的重复率是测试原文得到重复率的 1.6 倍，较大幅度地超出了原文重复率。而其他网站在此情况下测得的文章重复率接近测试原文得到的重复率，因此中国搜文章照妖镜比其他网站对改排版的操作要更敏感。将文章只改排版投入中国搜文章照妖镜网站的方式，查重结果的可信度比较低。此外，将文章只改公式时，大雅标准化后的数值为 0.57，即只改公式后大雅测得的重复率接近测试原文得到重复率的 1/2。其他网站在此情况下测得的文章重复率较接近测试原文得到的重复率，因此大雅网站比其他网站对只改公式操作要更敏感。所以，将文章只改公式投入大雅网站，查重结果可信度较低。

从一个网站对不同原文改动方式测得的结果来看，大雅网站对文章改动后测得的重复率普遍偏低，存在查重不够严谨的可能，因此大雅网站查重可信度偏低。而格子达网站对文章改动后测得的重复率普遍较高，因此格子达网站查重可信度有待后续考虑。从图 5 发现的规律如下：a）同一篇文章在不同的查重系统中检测出的重复率不同，甚至有的相差较大；b)改公式、改文字和改排版等方式都会改变查重结果；c)无意义地改动文字对降低重复率作用最大，只改排版对降低重复率作用最小；d)完全采用查重网站等机械工具来判断一篇论文是否抄袭，有时候是不准确的。因此，当同一篇论文在不同网站测得的查重结果相差较大时，有必要以人为判断做辅助。

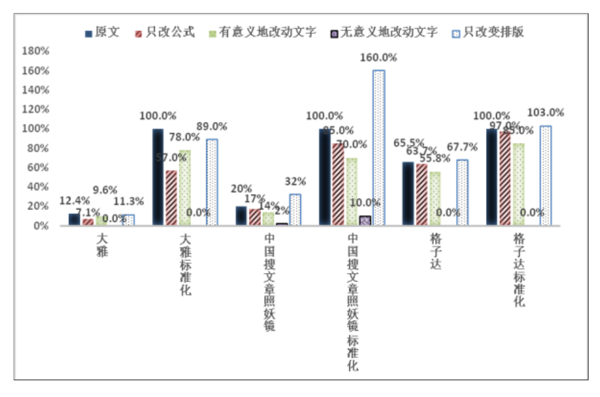


图 5 查重网站标准化检测

图 6 是根据格子达论文检测报告中的自写率、复写率和引用率这三个检测指标列出的。它是在不同原文改动方式下测得的自写率、复写率和引用率柱状图。其中，自写率即用全文有效片段总数 M 减去相似片段数 N(相似片段包括已作引用标示的内容)，然后除以全文有效片段总数 M 得到文章的自写率，公式为(M－N)/M。复写率即全文相似部分片段总字数 N 减去引用片段总字数 C，然后除以全文有效片段总字数 M，公式为(N－C)/ M。引用率即全文已加引用标示的片段总字数 C 除以有效片段总字数 M，公式 C/M。

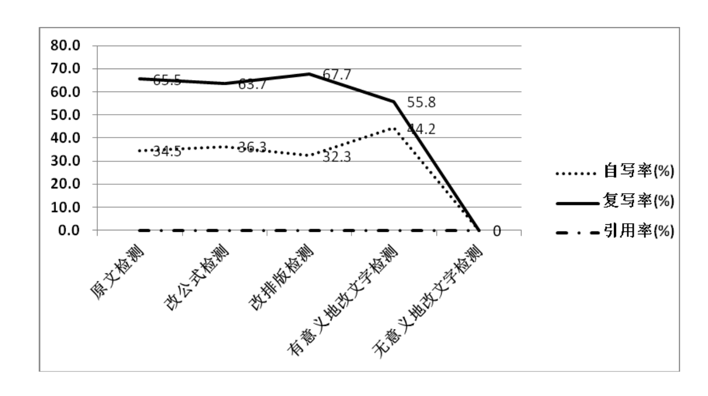


图 6 查重检测

从图 6 可以发现，使用格子达网站检测不同方式更改后文章的重复率，无意义地改文字检测到的自写率最高为 100%；复写率最低为 0。该种情况较可能使得论文真的抄袭没被找出，该漏洞的问题将留待未来的研究课题。

4 结束语

本文提出一种量化查重的方法，基于防止误判的论文发生，以补充互联网查重方式的不足。明显地，经由一些设计过的查重文章对部分网站查重做测试，整理出一些可以预期的论文问题比较其结果。由于各个查重网站的数据库以及检测算法的不同，导致同一篇论文在不同网站测得的查重结果可能出现较大差异。此时，完全采用查重网站等自动检测工具来判断一篇论文是否抄袭是不准确的。有必要加入人为判断，以降低误判因素的影响。

参考文献：

[1] Isoc D. Preventing plagiarism in engineering education and research [C]// Proc of International Symposium on FundAmentals of Electrical Engineering. 2014: 1-7.

[2] https://baike. baidu. com/item/%E4%BA%92%E8%81%94%E7%BD%91/199186?fr=aladdin. (Internet [DB/OL].

[3] 郭平，王可，罗阿理，等. 大数据分析中的计算智能研究现状与展望[J]. 软件学报, 2015, 26 (11): 3010-3025. (Guo Ping, Wang Ke, Luo Ali, *et al.* Present situation and prospect of computational intelligence in big data analysis [J]. Journal of Software, 2015, 26 (11): 3010-3025.)

[4] Irwanto M R, Zamara S B, Herdianto R, *et al.* SIPOC business model process to prevent plagiarism in an electronic journal [C]// Proc of the 3rd International Conference on Science in Information Technology. 2017: 492-497.

[5] 大雅论文查重网站 [EB/OL]. http://dsa. dayainfo. com/.

[6] 中国搜文章照妖镜论文查重网站 [EB/OL]. http://www. zhongguosou. com/zonghe/fanchaoxi. html.

[7] 格子达论文查重网站 [EB/OL]. http://www. gezida. com/.

[8] 知网 [EB/OL]. http://lwjc. cmscoo. cn/cncnki/?200.

[9] 维普网 [EB/OL]. http://gocheck. cn.

[10] 万方检测网站 [EB/OL]. http://www. wanfangdata. com. cn/. (Wanfangdata [DB/OL].

[11] PaperPass 论文查重网站 [EB/OL]. http://www. paperpass. com/.

[12] 知识产权卫士-拷克网 [EB/OL]. http://www. copycheck. com. cn/index. html.

[13] PaperFree 论文查重网站 [EB/OL]. http://www. paperfree. cn/.

[14] 论文狗 [EB/OL]. http://www. lunwengo. net/.

[15] PaperTest 论文查重网站 [EB/OL]. http://www. papertest. com/.

[16] Abdi A, Shamsuddin S M, Idris N, *et al.* A linguistic treatment for automatic external plagiarism detection [J]. Knowledge-Based Systems, 2017, 135 (11): 135-146.

[17] Gaspar P V, Velásquez Juan D. Docode 5: building a real-world plagiarism detection system [J]. Engineering Applications of Artificial Intelligence, 2017, 64 (9): 261-271.

[18] Markus Eckerstorfer, Eirik Malnes. Manual detection of snow avalanche debris using high-resolution Radarsat-2 SAR images [J]. Cold Regions Science and Technology, 2015, 120 (12): 205-218.

[19] 发表期刊查重率多少算合格——中国鸣网 [DB/OL]. http://mingmw. com/baike/bk/31523. html. (The publication of the journal weight rate is qualified-mingmw [EB/OL]. http://mingmw. com/baike/bk/31523. html.)

[20] 丁正生. 概率论与数理统计简明教程 [M]. 北京: 高等教育出版社, 2005: 125-127. (Ding Zhengsheng. Brief tutorial on probability theory and mathematical statistics [M]. Beijing: advanced education press, 2005: 125-127.)

[21] 徐洪文. 关于置信度选取问题的讨论 [EB/OL]. http://www. docin. com/p-729685972. html. (Xu Hongwen. Discussion on the selection of confidence degree. [EB/OL]. http://www. docin. com/p-729685972. html.)

[22] 孔欣欣, 苏本昌, 王宏志, 等. 基于标签权重评分的推荐模型及算法研究 [J]. 计算机学报, 2017, (6): 1440-1452. (Kong Xinxin, Su Benchang, Wang Hongzhi, *et al.* Recommendation model and algorithm based on label weight score [J]. Acta computer science, 2017, (6): 1440-1452.)

[23] 帮助高中生渡过数学学习困难期的几点尝试——论文网 [EB/OL]. http://www. lunwendata. com/thesis/2017/105980. html. (A few attempts to help high school students get through their math difficulties [EB/OL].)